Method and System for Separating Multiple Sound Sources from Monophonic Input with Non-Negative Matrix Factor Deconvolution

Field of the Invention

[01] The invention relates generally to the field of signal processing and in particular to detecting and separating components of time series signals acquired from multiple sources via a single channel.

[02] Background of the Invention

[03] Non-negative matrix factorization (NMF) has been described as a positive matrix factorization, see Paatero, "Least Squares Formulation of Robust Non-Negative Factor Analysis," Chemometrics and Intelligent Laboratory Systems 37, pp. 23-35, 1997. Since its inception, NMF has been applied successfully in a variety of applications, despite a less than rigorous statistical underpinning.

[04] Lee, et al, in "Learning the parts of objects by non-negative matrix factorization," Nature, Volume 401, pp.788-791, 1999, describe NMF as an alternative technique for dimensionality reduction. There, non-negativity constraints are enforced during matrix construction in order to determine parts of human faces from a single image.

[05] However, that system is restricted within the spatial confines of a single image. That is, the signal is strictly stationary. It is desired to extend

NMF for time series data streams. Then, it would be possible to apply NMF to the problem of source separation for single channel inputs.

[06] Non-Negative Matrix Factorization

[07] The conventional formulation of NMF is defined as follows. Starting with a complex non-negative M x N matrix $\mathbf{V} \in \Re^{\geq 0, MxN}$, the goal is to approximate the matrix $\mathbf{V}$ as a product of two simple non-negative matrices $\mathbf{W} \in \Re^{\geq 0, MxR}$ and $\mathbf{H} \in \Re^{\geq 0, RxN}$, where $R \leq M$, and an error is minimized when the matrix $\mathbf{V}$ is reconstructed approximately by $\mathbf{W} \bullet \mathbf{H}$.

[08] The error of the reconstruction can be measured using a variety of cost functions. Lee et al., use a cost function:

$$D = \left\| \mathbf{V} \otimes \ln\left(\frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}}\right) - \mathbf{V} + \mathbf{W} \cdot \mathbf{H} \right\|_F, \qquad (1)$$

where $\|.\|_F$ is the Frobenius norm, and $\otimes$ is the Hadamard product, i.e., an element-wise multiplication. The division is also element-wise.

[09] Lee et al., in "Algorithms for Non-Negative Matrix Factorization," Neural Information Processing Systems 2000, pp. 556-562, 2000, describe an efficient multiplicative update process for optimizing the cost function without a need for constraints to enforce non-negativity:

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}^\top \cdot \frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}}}{\mathbf{W}^\top \cdot \mathbf{1}}, \quad \mathbf{W} = \mathbf{W} \otimes \frac{\frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}} \cdot \mathbf{H}^\top}{\mathbf{1} \cdot \mathbf{H}^\top}, \qquad (2)$$

where 1 is an M x N matrix with all its elements set to unity, and the divisions are again element-wise. The variable R corresponds to the number

of basis functions to extract. The variable R is usually set to a small number so that the NMF results into a low-rank approximation.

[010] NMF for Sound Object Extraction

[011] It has been shown that sequentially applying principle component analysis (PCA) and independent component analysis (ICA) on magnitude short-time spectra results in decompositions that enable the extraction of multiple sounds from single-channel inputs, see Casey et al., "Separation of Mixed Audio Sources by Independent Subspace Analysis," Proceedings of the International Computer Music Conference, August, 2000, and Smaragdis, "Redundancy Reduction for Computational Audition, a Unifying Approach," Doctoral Dissertation, MAS Dept., Massachusetts Institute of Technology, Cambridge MA, USA, 2001.

[012] It is desired to provide a similar formulation using NMF.

[013] Consider a sound scene $s(t)$, and its short-time Fourier transform arranged into an M x N matrix:

$$\mathbf{F} = DFT\begin{bmatrix} s(t_1) & s(t_2) & & s(t_N) \\ \vdots & \vdots & \cdots & \vdots \\ s(t_1+M-1) & s(t_2+M-1) & & s(t_N+M-1) \end{bmatrix}, \qquad (3)$$

where M is a size of the discrete Fourier transform (DFT), and N is a total number of frames processed. Ideally, some window function is applied to the input sound signal to improve the spectral estimation. However, because the window function is not a crucial addition, it is omitted for notational simplicity.

[014] From the matrix $\mathbf{F} \in \Re^{MxR}$, the magnitude of the transform $\mathbf{V} = |\mathbf{F}|$, i.e., $\mathbf{V} \in \Re^{\geq 0, MxR}$ can be extracted, and then, the NMF can be applied.

[015] To better understand this operation, consider the plots 100 of a spectrogram 101, spectral bases 102 and corresponding time weights 103 in Figure 1. The plot 101 on the lower right is the input magnitude spectrogram. The plot 101 represents two sinusoidal signals with randomly gated amplitudes. Note, that the signals are from a *single* source, or monophonic signal .

[016] The two columns of the matrix **W** 102, interpreted as spectral bases, are shown in the lower left. The rows of **H** 103, depicted in the top, are the time weights corresponding to the two spectral bases of the matrix **W**. There is one row of weights for each column of bases.

[017] It can be seen that this spectrogram defines an acoustic scene that is composed of sinusoids of two frequencies 'beeping' in and out in some random manner. By applying a two-component NMF to this signal, the two factors **W** and **H** can be obtained as shown in Figure 1.

[018] The two columns of **W**, shown in the lower left plot 102, only have energy at the two frequencies that are present in the input spectrogram 101. These two columns can be interpreted as basis functions for the spectra contained in the spectrogram.

[019] Likewise the rows of **H**, shown in the top plot 103, only have energy at the time points where the two sinusoids have energy. The rows of **H** can be interpreted as the weights of the spectral bases at each time instance. The bases and the weights have a one-to-one correspondence. The first basis describes the spectrum of one of the sinusoids, and the first weight vector describes the time envelope of the spectrum. Likewise, the second sinusoid is described in both time and frequency by the second bases and second weight vector.

[020] In effect, the spectrogram of Figure 1 provides a rudimentary description of the input sound scene. Although the example in Figure 1 is simplistic, the general method is powerful enough to dissect even a piece of complex piano music to a set of weights and spectral bases describing each note played and its position in time for that note, effectively performing musical transcription, see Smaragdis et al., "Non-Negative Matrix Factorization for Polyphonic Music Transcription," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2003, and U.S. Patent Application 10/626,456, filed on July 23, 2003, titled "Method and System for Detecting and Temporally Relating Components in Non-Stationary Signals," incorporated herein by reference.

[021] The above described method works well for many audio tasks. However, that method does not take into account relative positions of each spectrum, thereby discarding temporal information. Therefore, it is desired to extend the conventional NMF so that it can be applied to multiple time series data streams so that source separation is possible from single channel input signals.

Summary of the Invention

[022] The invention provides a non-negative matrix factor deconvolution (NMFD) that can identify signal components with a temporal structure. The method and system according to the invention can be applied to a magnitude spectrum domain to extract multiple sound objects from a single channel auditory scene.

[023] A method and system separates components in individual signals, such as time series data streams.

[024] A single sensor acquires concurrently multiple individual signals. Each individual signal is generated by a different source.

[025] An input non-negative matrix representing the individual signals is constructed. The columns of the input non-negative matrix represent features of the individual signals at different instances in time.

[026] The input non-negative matrix is factored into a set of non-negative bases matrices and a non-negative weight matrix. The set of bases matrices and the weight matrix represent the plurality of individual signals at the different instances of time.

Brief Description of the Drawings

[027] Figure 1 are plots of a spectrogram, bases and weights of a non-negative matrix factorization of a sound scene according to the prior art;

[028] Figure 2 are plots of a spectrogram, bases and weights of a non-negative matrix factor deconvolution of a sound scene according to the invention;

[029] Figure 3 are plots of a spectrogram, bases and weights of a non-negative matrix factor deconvolution of a sound scene according to the invention; and

[030] Figure 4 is a block diagram of a system and method according to the invention.

Detailed Description of the Preferred Embodiment

[031] Non-Negative Matrix Factor Deconvolution

[032] The invention provides a method and system that uses a non-negative matrix factor deconvolution (NMFD). Here, deconvolving means 'unrolling' a complex mixture of time series data streams into separate elements. The invention takes into account relative positions of each spectrum in a complex input signal from a single channel. This way multiple

signal sources of time series data streams can be separated from a single input channel.

[033] In the prior art, the model used is $\mathbf{V} \approx \mathbf{W} \bullet \mathbf{H}$. The invention extends this model to:

$$\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overrightarrow{\mathbf{H}}^{t} \quad , \tag{4}$$

where an input matrix $\mathbf{V} \in \Re^{\geq 0, MxN}$ is decomposed to a set of non-negative bases matrices $\mathbf{W}_t \in \Re^{\geq 0, MxR}$ and a non-negative weight matrix $\mathbf{H} \in \Re^{\geq 0, RxN}$, over successive time intervals. The operator $\overrightarrow{(.)}^{t}$ shifts the columns of the matrix $\mathbf{H}$ by i time increments to the right, for example

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}, \overrightarrow{\mathbf{A}}^{0} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}, \overrightarrow{\mathbf{A}}^{1} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{bmatrix}, \overrightarrow{\mathbf{A}}^{2} = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 5 & 6 \end{bmatrix}, \cdots \tag{5}$$

[034] The left most columns of the matrix $\mathbf{H}$ are appropriately set to zero to maintain the original size of the input matrix. Likewise, an inverse operation $\overleftarrow{(.)}^{i}$ shifts columns of the weight matrix $\mathbf{H}$ to the left by i time increments.

[035] The objective is to determine sets of bases matrices $\mathbf{W}_t$ and the weight matrix $\mathbf{H}$ to approximate the input matrix $\mathbf{V}$ representing the input signal as best as possible.

[036] Cost Function to Measure Error of Reconstruction

[037] A value $\mathbf{\Lambda}$ is set $\sum_{t=0}^{T-1}\mathbf{W}_t \bullet \overrightarrow{\mathbf{H}}$, and a cost function to measure an error of the reconstruction is defined as

$$D = \left\| \mathbf{V} \otimes \ln(\frac{\mathbf{V}}{\mathbf{\Lambda}}) - \mathbf{V} + \mathbf{\Lambda} \right\|_F . \quad (6)$$

[038] In contrast with the prior art, where $\mathbf{\Lambda} = \mathbf{W} \bullet \mathbf{H}$, using a similar notation, the invention has to optimize more than two matrices over multiple time intervals to optimize the cost function.

[039] To update the cost function for each iteration of t, the columns are shifted to appropriately line up the arguments according to:

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}_t^\top \cdot [\frac{\mathbf{V}}{\mathbf{\Lambda}}]^{\leftarrow t}}{\mathbf{W}_t^\top \cdot \mathbf{1}} \text{ and } \mathbf{W}_t = \mathbf{W}_t \otimes \frac{\frac{\mathbf{V}}{\mathbf{\Lambda}} \cdot \overrightarrow{\mathbf{H}}^{t\top}}{\mathbf{1} \cdot \overrightarrow{\mathbf{H}}^{t\top}}, \ \forall t \in [0 \ldots T-1] . \quad (7)$$

[040] In every iteration for each time interval t, the matrix $\mathbf{H}$ and each matrix $\mathbf{W}_t$ is updated. That way, the factors can be updated in parallel and account for their interaction. In complex cases it is often useful to average the updates of the matrix $\mathbf{H}$ over all time intervals t. Due to the rapid convergence properties of the multiplicative rules, there is the danger that the matrix $\mathbf{H}$ is influenced by the previous matrix $\mathbf{W}_t$ used for its update, rather than the entire set of matrices $\mathbf{W}_t$.

[041] Example Deconvolution

[042] To gain some intuition on the form of the factors $\mathbf{W}_t$ and $\mathbf{H}$, consider the plots in Figure 2, which shows and extracted NMFD bases and weights. The lower right plot 201 is a magnitude spectrogram that is used as an input to NMFD method according to the invention. Note, that signals vary over time, are generated by multiple sources, and are acquired via a single channel.

[043] The two lower left plots 202 are derived from the factors $\mathbf{W}_t$, and are interpreted as temporal-spectral bases. The rows of the factor $\mathbf{H}$, depicted at the top plot 203, are the time weights corresponding to the two temporal-spectral bases. Note that the lower left plot 202 has been zero-padded from left and right so as to appear in the same scale as the input plot.

[044] Like the example shown for the scene shown in Figure 1, the spectrogram contains two randomly repeating elements, however, in this case, the elements exhibit a temporal structure, which cannot be expressed by spectral bases spanning a single time interval, as in the prior art.

[045] A two-component NMFD with T = 10 is applied. This results into a factor $\mathbf{H}$ and $T \times \mathbf{W}_t$ matrices of size $M \times 2$. The n^{th} column of the t^{th} $\mathbf{W}_t$ matrix is the n^{th} basis offset by t increments in the left-to-right dimension, time in this case. In other words, the $\mathbf{W}_t$ matrices contain bases that extend in both dimensions of the input. The factor $\mathbf{H}$, like the conventional NMF, holds the weights of these functions. Examining Figure 2, it can be seen that

the bases in the set of factors $\mathbf{W}_t$ contain the finer temporal information in the sound patterns, while the factor **H** localizes the patterns in time.

[046] **NMFD for Sound Object Extraction**

[047] Using the above formulation of NMFD, a sound segment, which contains a set of drum sounds, can be analyzed. In this example, the drum sounds exhibit some overlap in both time and frequency. The input is sampled at 11.025 Hz and analyzed with 256-point DFTs with an overlap of 128-points. A Hamming window is applied to the input to improve the spectral estimate. The NMFD is performed for three basis functions, each with a time extend of ten DFT frames, i.e., $R = 3$ and $T = 10$.

[048] Figure 3 shows the spectrogram plot 301, and the corresponding bases and weight factor plots 302-303 for the scene, as before. There are three types of drum sounds present into the scene including four instances of a bass drum sound at low frequencies, two instances of a snare drum sound with two loud wideband bursts, and a 'hi-hat' drum sound with a repeating high-band burst.

[049] The lower right plot 301 is the magnitude spectrogram for the input signal. The three lower left plots 302 are the temporal-spectral bases for the factors $\mathbf{W}_t$. Their corresponding weights, which are rows of the factor **H,** are depicted at the top plot 303. Note how the extracted bases encapsulate the temporal/spectral structure of the three drum sounds in the spectrogram 301.

[050] Upon analysis, a set of spectral/temporal basis functions are extracted from $\mathbf{W}_t$. The weights from the factor $\mathbf{H}$ show when these bases are placed in time. The bases encapsulated the short-time spectral evolution of each different type of drum sound. For example, the second basis (2) adapts to the bass drum sound structure. Note how the main frequency of this basis decreases over time and is preceded by a wide-band element just like the bass drum sound. Likewise the snare drum basis (3) is wide-band with denser energy at the mid-frequencies, and the hi-hat drum basis (1) is mostly high-band sound.

[051] A reconstruction can be performed to recover the full spectrogram or partial spectrograms for any one of the three input sounds to perform source separation. The partial reconstruction of the input spectrogram is performed using one basis function at a time. For example, to extract the bass drum, which was mapped to the j^{th} basis perform:

$$\hat{\mathbf{V}}_j = \sum_{t=0}^{T-1} \mathbf{W}_t^{(j)} \cdot \overset{t\rightarrow}{\mathbf{H}} \quad , \tag{8}$$

where the $(\overset{t\rightarrow}{.})^{(j)}$ operator selects the j^{th} column of the argument. This yields an output non-negative matrix representing a magnitude spectrogram of just one component of the input signal. This can be applied to original phase of the spectrogram. Inverting the result yields a time series of just, for example, the base drum sound.

[052] Subjectively, the extracted elements consistently sound substantially like the corresponding elements of the input sound scene. That is, the reconstructed base drum sound is like the base drum sound in the

input mixture. However, it is very difficult to provide a useful and intuitive quantitative measure that otherwise describes the quality of separation due to various non-linear distortions and lost information, problems inherent in the mixing and the analysis processes.

System Structure and Method

[053] As shown in Figure 4, the invention provides a system and method for detecting components of non-stationary, individual signals from multiple sources acquired via a single channel, and determining a temporal relationship among the components of the signals.

[054] The system 400 includes a sensor 410, e.g., microphone, an analog-to digital (A/D) converter 420, a sample buffer 430, a transform 440, a matrix buffer 450, and a deconvolution factorer 500, serially connected to each other.

[055] Multiple acoustic signals 401 are generated concurrently by multiple signal sources 402, for example, three different types of drums. The sensor acquires the signals concurrently. The analog signals 411 are provided by the single sensor 410, and converted 420 to digital samples 421 for the sample buffer 430. The samples are windowed to produce frames 431 for the transform 440, which outputs features 441, e.g., magnitude spectra, to the matrix buffer 450. An input non-negative matrix V 451 representing the magnitude spectra is deconvolutionally factored 500 according to the invention. The factors W*t* 510 and H 520 are respectively bases and weights that represent a separation of the multiple acoustic signals 401. A

reconstruction 530 can be performed to recover the full spectrogram 451 or partial spectrograms 531-533, i.e., each an output non-negative matrix, for any one of the three input sounds. The output matrices 531-533 can be used to perform source separation 540.

Effect of the Invention

[056] The invention provides a convolutional non-negative matrix factorization. version of NMF that overcomes the problems with the conventional NMF when analyzing temporal patterns. This extension results in an extraction of more expressive basis functions. These basis functions can be used on spectrograms to extract separate sound sources from a sound scenes acquired by a single channel, e.g., one microphone.

[057] Although the example application used to describe the invention uses acoustic signals, it should be understood that the invention can be applied to any time series data stream, i.e., individual signals that were generated by multiple signal sources and acquired via a single input channel, e.g., sonar, ultrasound, seismic, physiological, radio, radar, light and other electrical and electro-magnetic signals.

[058] Although the invention has been described by way of examples of preferred embodiments, it is to be understood that various other adaptations and modifications may be made within the spirit and scope of the invention. Therefore, it is the object of the appended claims to cover all such variations and modifications as come within the true spirit and scope of the invention.